DOES PHYSICAL INTERPRETABILITY OF OBSERVATION MAP IMPROVE THE PHOTOMETRIC STEREO NETWORK?

Satoshi Ikehata[1]

[1]National Institute of Informatics, Tokyo, Japan

ABSTRACT

In this paper, we revisit *observation map* which is the input representation for the deep photometric stereo networks where pixelwise observations under different lights are protectively integrated to handle an arbitrary number of input images. Based on the hypothesis that the physical interpretability of observation map contributes to its performance, we empirically validate it by proposing two novel ideas; one is a pixelwise unified inverse rendering framework which accounts the physical reasoning to recover the surface normals and the other is the network architecture that is equivariant/invariant to the view-axis-around rotation of the pixelwise observation map. By introducing these two ideas, our experimental evaluation on the public dataset indicated that more explicit physical reasoning of observation map improves the performance of the photometric stereo task.

Index Terms— photometric stereo, observation map

1. INTRODUCTION

Photometric stereo [1] is a widely researched task for decades which aims at recovering the surface normal map of an object from images captured under different lights with a fixed camera. Since the classical physics-based photometric stereo algorithms [2, 3, 4] were hardly applicable to objects with complex non-convex geometry and non-Lambertian reflectance, recent state-of-the-art photometric stereo algorithms adopt the data-driven approach, using the deep neural networks [5, 6, 7, 8]. Unlike other computer vision tasks, the photometric stereo networks must accept an arbitrary number of images as input and pursuing the proper data aggregation strategy has been a major interest in recent studies [7, 9, 10, 11, 12]. At current time, one of the most promising strategy is based on *observation map* [8] due to its simplicity and effectiveness.

Observation map is a 2-d matrix where pixelwise observations under arbitrary number of directional lights are integrated. As shown in Fig. 1, each matrix cell corresponds to a discretized light direction on a unit hemisphere, and each observation (*i.e.*, pixel color) is projected to a point on the map according to its light direction. In the photometric stereo task, an observation map is individually encoded at each pixel and fed into the neural networks to predict its surface normal.

Though photometric stereo algorithms based on observation map [8, 13, 14, 15] have demonstrated the state-of-the-art performance on the public benchmark [16], all the current algorithms have just let networks memorize the patterns of an observation map and surface normal and there is no clear solution when no identical pattern matches exist due to the huge possible combinations of geometries, materials and lightings.

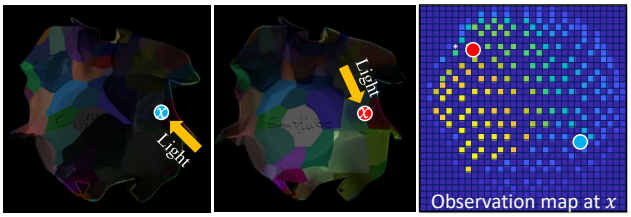

Fig. 1. Example of a physically interpretable observation map at a nonconvex surface point, which reasonably encodes its surface material and geometry.

As for this point, we pay attention to the *physical interpretability* of observation map, which means that an observation map tells us the attributes at around a surface point. For instance, Fig. 1 shows an observation map at a surface point whose surface normal is pointing down left. The radially decreasing intensities clues that the light is reflected on the rough dielectric surface (material). On the other hand, the abrupt change of values at right side evidences the presence of cast shadows therefore the surface is non-convex (geometry). Interpreted observation map in this way, we can also infer the surface normal direction as a part of underling Bidirectional Reflectance Distribution Function (BRDF) [17] which is visible in observation map as a reflectance lobe. This physical interpretability is definitely an advantage of observation map, not in other strategies such as set-pooling [7, 9], graph-convolution [10] and self-attention [11, 12] in recent literature. Nevertheless, there were no studies that have examined the effects of physical interpretability of observation map.

The goal of this work is to take advantage of physical interpretabilty of observation map more explicitly to use it more effectively. Following this motivation, we propose two ideas to improve the photometric stereo networks using observation map. First, we integrate the unsupervised inverse rendering framework into the näive regression network. We ask neural networks to parse the observation map into the physical intrinsic attributes (*e.g.*, surface normal, surface roughness, surface base color) and to integrate them in a physically plausible manner with the inverse rendering loss. Second, we explicitly account the reflectance isotropy [18], which implies that the surface normal prediction should be equivariant to the rotation of the observation map. In the prior work [8], this property was enforced based on the external rotational data augmentation. However, we will verify that the internal encouragement of the isotropy in neural networks also boost the

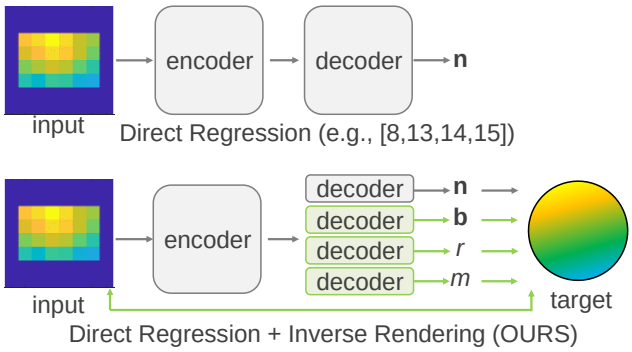

Fig. 2. We integrate the inverse rendering pipeline into the conventional Direct Regression framework for enhancing the physical interpretability of observation map.

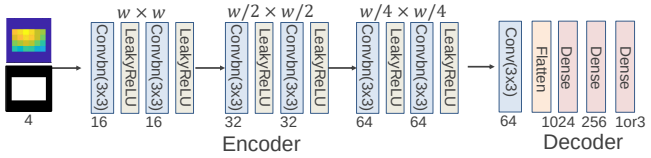

Fig. 3. The architectures of our encoder and decoder.

performance even without using any data augmentation.

We validate our ideas on the DiLiGenT Photometric Stereo Dataset [16] as well as compare them against recent photometric stereo algorithms [6, 7, 8] and show that the neural networks can also successfully interpret observation map in a physically plausible manner.

2. PRELIMINARIES

Problem Formulation: Photometric stereo is a problem to recover the unit normal vector $\mathbf{n} \in \mathbb{R}^3$ for each pixel from a collection of observations $I_{1\leq j\leq m} \in \mathbb{R}^3$ under m different lighting directions $\mathbf{l}_{1\leq j\leq m} \in \mathbb{R}^3$. Under the calibrated setup, pixelwise observations are normalized by corresponding light intensities and the view direction is fixed by $\mathbf{v} = [0\ 0\ 1]$. Henceforth, we rely on the classical assumptions of fixed, linear orthographic camera and known directional lighting.

Observation Map: Observation map [8] is projections of appearances at a surface point onto a 2-D matrix based on their light directions ($\mathbf{l}_{1\leq j\leq m} \triangleq [l_x^j\ l_y^j\ l_z^j]^\top$). Each element of an observation map $O \in \mathbb{R}^{w\times w}$ is defined as follow:

$$O_{\mathrm{int}(w(l_x^j+1)/2),\mathrm{int}(w(l_y^j+1)/2)} = \alpha I_j. \tag{1}$$

Here "int" is an operator to round a floating value to an integer. The size of the observation map (w) and the choice of the scaling factor (α) are arbitrary but were suggested to be set by 32 and $1/\max(I)$, respectively in [8] and we follow them.

3. METHOD

Our goal is to take advantage of physical interpretability of observation map to improve the photometric stereo network taking it as input. First of all, we define the "baseline" network and extend it according to our ideas. As shown in Fig. 2, the simplest strategy to predict the surface normal from an observation map would be the "Direct Regression" where neural networks are trained only by the supervision of the surface normal. All the photometric stereo algorithms using observation map is based on this strategy [8, 13, 14, 15] with no physical reasoning behind. On the other hand, to explicitly utilize physical interpretability of observation map, we integrate multiple decoders to predict surface attributes (*e.g.*, roughness, base color, metalness) with the baseline architecture and synthetically render the observation map based on the predefined image formation model. Unsupervised *inverse rendering loss* unifies all the attributes to confirm that they are predicted in the physically plausible manner. We should note that if we focus only on the normal prediction, we have the decoder of same parameter size with the baseline, but we expect that explicit physical reasoning by the additive inverse rendering would improve the surface normal prediction.

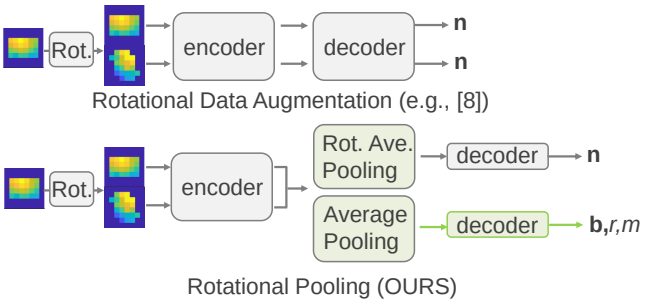

Fig. 4. Conventionally, the equivariance to rotation of observation map was realized by data augmentation [8]. On the other hand, we propose the pooling-based approach for realizing the rotation invariance/equivariance.

In addition, we explicitly account the reflectance isotropy and inherent equivariance/invariance of surface attributes to rotation of observation map. As illustrated in Fig. 4, instead of applying data augmentation as in existing works [8, 15], we propose the pooling-based strategy which merges all the feature maps from differently rotated observation maps so that the prediction of surface normal is approximately equivariant, and one of other surface attributes is invariant to the rotation of observation map. We now describe details.

Inverse Rendering of Observation Map: We here detail the physical formation model of observation map for our inverse rendering pipeline. We use a simplified version of Principled BRDF [19] which has been commonly used for synthesizing photometric stereo datasets [8, 12]. Assuming that we don't consider the anisotropic reflection, subsurface scattering, sheen and clear coat which all are not dominant in real-world isotropic materials, our BRDF (f) is controlled by three parameters, *base color* ($\mathbf{b} \in \mathbb{R}^3$), *roughness* ($r \in [0, 1]$) and *metalness* ($m \in [0, 1]$) besides the surface normal and lighting. Please refer to [19] for mathematical descriptions, however we note that the specular lobe in Principled BRDF is the Cook-Torrance microfacet BRDF model using the GGX distribution as the microfacet normal distribution. Based on this BRDF, the observation map ($\hat{O}$) is numerically

Table 1. The ablation study of the rotation pooling.

	Ba	Be	Bu	Ca	Co	Go	Ha	Po1	Po2	Re	Ave
K (Train) = 20											
K (Test) = 1	10.4	13.0	16.4	17.3	18.2	16.7	27.9	12.6	14.0	21.3	16.8
K (Test) = 5	3.0	4.7	8.0	5.0	5.7	7.5	14.7	5.2	5.8	11.4	7.1
K (Test) = 10	2.6	4.0	7.8	4.5	5.5	7.2	14.2	**4.9**	5.5	**10.4**	6.7
K (Test) = 20	2.5	4.0	7.6	4.5	5.4	6.9	14.1	5.0	**5.3**	10.5	6.6
K (Test) = 40	**2.5**	**3.9**	**7.6**	**4.5**	5.4	**6.9**	14.0	5.0	**5.3**	10.4	**6.6**
K (Test) = 90	2.6	**3.9**	7.6	4.5	**5.4**	6.9	**14.0**	5.0	5.3	10.6	6.6
K (Train) = 10											
K (Test) = 1	7.6	10.0	12.6	10.2	12.8	9.1	17.9	10.4	9.7	15.0	11.5
K (Test) = 5	2.6	4.2	8.0	4.5	6.3	7.4	13.9	5.2	5.5	11.1	6.9
K (Test) = 10	**2.3**	3.9	7.7	**4.2**	5.7	7.2	**13.8**	**5.0**	5.4	**10.7**	**6.6**
K (Test) = 20	2.3	3.9	**7.7**	4.3	5.6	**7.1**	13.8	5.0	**5.4**	11.4	6.6
K (Test) = 40	2.4	**3.9**	**7.7**	4.3	**5.5**	7.2	13.8	5.0	5.4	10.7	6.6
K (Train) = 1											
K (Test) = 1	**2.8**	**4.1**	**8.0**	**4.6**	**7.1**	**7.6**	**14.0**	**5.3**	**5.7**	**11.4**	**7.1**
K (Test) = 5	13.2	15.2	16.1	12.3	16.9	16.9	20.1	14.0	15.0	21.3	16.1
K (Test) = 10	15.5	16.9	17.8	13.9	18.6	18.7	21.7	15.6	16.7	22.6	17.8
K (Test) = 20	16.5	17.7	18.6	14.6	19.4	19.6	22.6	16.2	17.7	23.5	18.6
K (Test) = 40	17.2	18.3	19.1	15.1	19.8	20.1	23.1	16.7	18.3	23.9	19.2

formulated as follow:

$$\hat{O}(\mathbf{l}) = \max\{\mathbf{n}^T\mathbf{l} * f(\mathbf{n}, \mathbf{b}, r, m; \mathbf{l}) * S(\theta), 0\} \quad \forall\, \mathbf{l}, \tag{2}$$

where $\max$ operator accounts for the attached shadow and $S(\theta)$ is named *observation mask* which is actually a sampling operator to sample light directions of *input* observation map.

Consideration of Reflectance Isotropy: As has been detailed in [8], the surface reflectance isotropy implies that the surface normal prediction should be *equivariant* to the rotation of observation map, which is formally described as $r(g(x)) = g(r(x))$ where x is the input observation map, g is the network and r is the rotation matrix which rotates the lighting and surface normal directions $(\mathbf{n}, \mathbf{l})$ around the viewing direction. Similarly, predictions of other attributes $(\mathbf{b}, r, m)$ are rotation *invariant* as $g(x) = g(r(x))$. We can intuitively confirm them by rotating the actual observation map in Fig. 1 and seeing the reflectance lobe.

Though rotational data augmentation in [8] had improved the surface normal prediction, it was unclear if the improvement had really come from the care of reflectance isotropy rather than from the larger training sample size. To clarify, we propose the rotation equivariant/invariant networks without relying on data augmentation but on *rotational pooling*. As illustrated in Fig. 4, copies of input observation map are rotated at regular intervals (*i.e.*, $360/K$, K is the number of rotations) and each copy is fed to the same encoder. The rotation equivariance implies that the feature map from a rotated observation map should also be rotated accordingly, in other words, feature maps rotated backwards should be close with each other. Based on this logic, we inversely rotate and average feature maps of different rotation angles and pass the result to the surface normal decoder. We call this process as *rotated average pooling*. Since the pooling operation is an order-agnostic operation, if all the features rotated by from 0 to 360 degrees are pooled together, theoretically the original rotation angle of observation map doesn't affect the output. Similarly, when decoding surface material attributes, we apply the basic *average pooling* operation on all the encoded feature maps and then feed the pooled feature to decoders to encourage them to output the same values regardless of the rotation angles of the input observation map (*i.e.*rotation invariance). By introducing these operations, we expect both encoder and decoder get constrained by reflectance isotropy without explicitly increasing training samples.

Table 2. The evaluation on *DiLiGenT* dataset with 96 images.

	Ba	Be	Bu	Ca	Co	Go	Ha	Po1	Po2	Re	Ave
D.R.+I.R.L.+R.P.	2.3	**3.9**	**7.7**	4.2	**5.7**	**7.2**	**13.8**	**5.0**	**5.4**	**10.7**	**6.6**
D.R.+R.P.	2.0	4.1	8.2	4.8	6.0	7.5	14.1	5.2	6.0	11.4	6.9
D.R.+I.R.L.	2.8	4.1	8.0	4.6	7.1	7.6	14.0	5.3	5.7	11.4	7.1
D.R.+R.A.	2.1	4.2	8.1	4.4	7.9	7.4	13.8	5.5	6.4	12.3	7.2
D.R.	2.6	4.7	8.7	**3.9**	8.1	7.3	14.2	5.9	6.5	12.6	7.5
PS-FCN [7]	2.8	7.6	7.9	6.2	7.3	8.6	15.9	7.1	7.6	13.3	8.4
Taniai [6]	**1.5**	5.8	10.4	5.4	6.3	11.5	22.6	6.1	7.8	11.0	8.8
Woodham [1]	4.1	8.4	14.9	8.4	25.6	18.5	30.6	8.9	14.7	19.8	15.4

Table 3. The evaluation on *DiLiGenT* dataset with 10 images.

	Ba	Be	Bu	Ca	Co	Go	Ha	Po1	Po2	Re	Ave
D.R.+I.R.L.+R.P.	4.3	5.4	**8.7**	**6.2**	11.6	10.7	20.6	7.0	8.0	**13.2**	**9.6**
D.R.+R.P.	5.3	**5.3**	9.2	6.3	13.5	11.0	18.6	7.4	7.9	14.4	9.9
D.R.+I.R.L.	5.2	6.2	10.4	7.1	12.7	12.1	19.5	**6.4**	**7.6**	15.3	10.2
D.R.+R.A.	9.6	14.8	15.0	11.7	15.0	16.4	21.5	12.7	15.9	16.2	14.9
D.R.	10.6	16.8	15.4	12.5	16.8	17.4	22.0	13.4	16.9	16.8	15.9
PS-FCN [7]	4.0	7.2	9.8	8.3	10.5	11.6	18.7	10.1	9.9	15.0	10.5
SPLINE-Net [13]	**4.0**	8.7	11.4	6.7	10.2	10.5	**17.3**	7.3	9.7	14.4	10.0
Minify-Net [14]	5.0	6.0	10.1	7.5	**8.8**	**10.4**	19.1	8.8	11.8	16.1	10.4

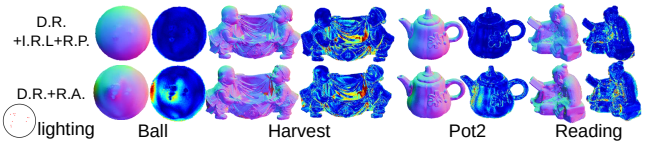

Fig. 5. The qualitative comparison on *DiLiGenT* dataset with 10 images. The lighting configuration is presented.

Network Architecture and Training Loss: To clarify our discussion point about the advantage of physical interpretability of observation map, we design the simple encoder and decoder. Please see details in Fig. 3 and we won't go into detail here due to the space limit.

Our training loss is comprised of the normal reconstruction loss $\mathcal{L}_N$ and the scale-invariant inverse rendering loss $\mathcal{L}_I$. The normal reconstruction loss $\mathcal{L}_N$ penalizes for the distance between the predicted and ground truth surface normals. In our experiment, we used the standard ℓ_2 distance for this loss function. The unsupervised, scale-invariant inverse rendering loss $\mathcal{L}_I$ penalizes for the distance between the input and reconstructed observation maps.

$$\mathcal{L}_I = \text{smooth}_{\ell_1}(O, \beta\hat{O}). \tag{3}$$

Here O is the input observation map and $\hat{O}$ is given by Eq. (2). $\text{smooth}_{\ell_1}(p, q)$ [20] is a scale-invariant ℓ_1 function and its scaling factor β is analytically computed by minimizing

$$\beta \leftarrow \text{argmin}_{\hat{\beta}} \|O - \hat{\beta}\hat{O}\|_2^2. \tag{4}$$

Note that the scale-invariant loss is required since there is the ambiguity of the intensity scale between O and $\hat{O}$.

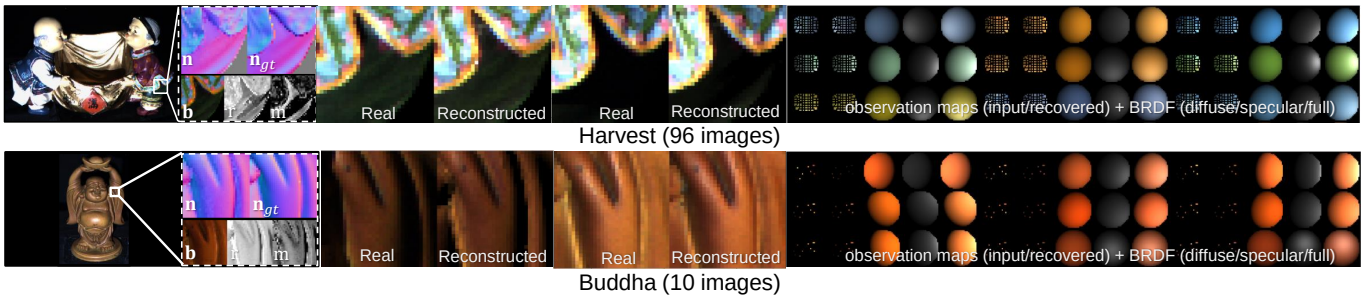

Fig. 6. The surface attribute reconstruction results by our D.R.+I.R.L.+R.P. architecture. Top: Harvest with 96 images. Bottom: Buddha with 10 images. From left to right: One of input images, recovered surface attributes, two reconstructed images under *novel* lightings, reconstructed observation maps (input, reconstructed, diffuse/specular BRDFs and full BRDF).

4. EXPERIMENTS

Training and Test Details: We followed the exactly same training strategy as [8] except that we trained our network using *CyclesPS+* dataset [12] which consists of 25 objects. Each object provides 32-bit floating images with a resolution of 256×256 under 740 different known lighting directions. Our network was trained on $3\times$ Nvidia Geforce GTX 1080Ti with Adam optimizer for 20 epochs, a batch size of 256 and a learning rate of 0.0002. We evaluate our method on *DiLiGenT* [16] which is a public benchmark dataset of 10 real objects (Ba:Ball, Be:Bear, Bu:Buddha, Ca:Cat, Co:Cow, Go:Goblet, Ha:Harvest, Po1:Pot1, Po2:Pot2, Re:Reading). Each object provides 16-bit integer images with a resolution of 612×512 from 96 different known lighting directions and the ground truth surface normal maps. For the evaluation, we simply compute the mean angular error (MAE) of predicted normal maps in degrees. Errors are averaged over 100 trials when the number of images is ten. In our experiment, we would like to show that baseline (*i.e.*, Direct Regression: D.R.) is enhanced with Inverse Rendering Loss (I.L.R.) and Rotational Pooling (R.P.) that are more physically interpretable architectures. Note that D.R. and D.R.+R.A. (*i.e.*, Rotational Augmentation) are equivalent to CNN-PS [8] (*i.e.*, w/o and w/ rotational augmentation) except that the network architecture was slightly simplified.

Analysis on Rotational Pooling: First, we investigate the effect of the number of rotation angles for the pooling in our D.R.+I.R.L+R.P architecture. Note that since there is no learnable parameters in R.P., the values at training and test could be different. As shown in Table 1, the optimal results are basically obtained when the number of rotations in training and test are close. Though there are multiple optimal choices, we choose $K(\text{Train}) = K(\text{Test}) = 10$ in following experiments simply Since they were also recommended in [8].

Quantitative Comparison on Dense Setup: We compared variants of our architectures on the dense setup (*i.e.*, all 96 images are used) to validate the effects of the inverse rendering loss and rotational pooling. Just for the reference, we also lined up some recent deep-learning-based algorithms of PS-FCN [7], Taniai and Maehara [6] and the conventional Lambertian method [1]. The results are illustrated in Table 2. Both I.R.L and R.P. enhanced the performance of D.R. as expected. Interestingly, R.P. showed the obvious advantage over R.A. which indicates that the R.P. could better consider the surface isotropy rather than R.A. due to the explicit enhancement of equivariance/invariance of the networks.

Quantitative Comparison on Sparse Setup: We also evaluated our method on sparse photometric stereo setup (*i.e.*, 10 images are used). Here we instead compared our method against Minify-Net[13] and SPLINE-Net [14] that are also based on observation map but tuned for the sparse setup, as well as PS-FCN [7]. The results are illustrated in Table 3 and Fig. 5. Though the tendency of the result is similar with one of the dense setup, the numerical improvement was much bigger. This could be due to the fact that the simple pattern matching in D.R. is more difficult in the sparse observation map, making physical reasoning more important.

Analysis on the Rendered Observation Map: Finally, we displayed the reconstructed surface attributes and rendering results of observation map in Fig. 6. Since there is no true label, we can only qualitatively discuss the result but it is obvious that the reconstructed observation map is almost identical to the input observation map which means that our inverse rendering pipeling behaved as intended. Since the recovered BRDF gives elements of observation map for *all the lighting directions*, it can be used to render the image under the novel lightings. The quality of synthesized image also shows that our method reasonably worked to capture the physics properties in observation map.

5. CONCLUSION

In this paper, we presented ideas to take advantages of the physical interpretability of observation map for the photometric stereo task. Our evaluation demonstrated that integrating the inverse rendering pipeline and the rotational pooling into the basic direct regression networks improve the performance without increasing the complexity of the network architecture nor using augmented training examples.

In the future work, we are interested in testing our ideas on more complicated photometric stereo networks that uses observation map as input (*e.g.*, [15]) to verify that this result is general. In addition, we also want to discuss the physical aspects in other aggregation strategies such as set-pooing [7] and self-attention [12] and discuss connections between data-driven approach and physics-based approach.

6. REFERENCES

[1] P. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Engg*, vol. 19, no. 1, pp. 139–144, 1980.

[2] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and spatially-varying brdfs from photometric stereo," *IEEE TPAMI*, vol. 32, no. 6, pp. 1060–1071, 2010.

[3] S. Ikehata and K. Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *CVPR*, 2014.

[4] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE TPAMI*, vol. 36, no. 6, pp. 1078–1091, 2014.

[5] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network," in *International Workshop on Physics Based Vision meets Deep Learning (PBDL) in Conjunction with IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] T. Taniai and T. Maehara, "Neural Inverse Rendering for General Reflectance Photometric Stereo," in *ICML*, 2018.

[7] G. Chen, K. Han, and K-Y. K. Wong, "Ps-fcn: A flexible learning framework for photometric stereo," *ECCV*, 2018.

[8] S. Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *ECCV*, 2018.

[9] Yakun Ju, Junyu Dong, and Sheng Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Transactions on Image Processing*, vol. 30, pp. 3676–3690, 2021.

[10] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "Gps-net: Graph-based photometric stereo network," *NeurIPS*, 2020.

[11] Huiyu Liu, Yunhui Yan, Kechen Song, and Han Yu, "Sps-net: Self-attention photometric stereo network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[12] S. Ikehata, "Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism," in *BMVC*, 2021.

[13] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *CVPR*, 2019.

[14] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L-Y. Duan, and A.C. Kot, "Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks," *ICCV*, 2019.

[15] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla, "Px-net: Simple and efficient pixel-wise training of photometric stereo networks," in *CVPR*, 2021, pp. 12757–12766.

[16] B. Shi, Z. Mo, Z. Wu, D.Duan, S-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE TPAMI*, p. (to appear), 2018.

[17] R. Montes and C. Urena, "An overview of brdf models," Tech. Rep., LSI-2012-001 en Digibug Coleccion: TIC167 - Articulos, 2012.

[18] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM TOG*, vol. 22, no. 3, pp. 759–769, 2003.

[19] B. Burley, "Physically-based shading at disney, part of practical physically based shading in film and game production," *SIGGRAPH 2012 Course Notes*, 2012.

[20] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.